Generalized Additive Models

Statistics 135

Autumn 2005

Generalized Additive Models

GAMs are one approach to non-parametric regression in the multiple predictor setting. The additive linear model is of the form

$$E[Y|X_1, \ldots, X_p] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

The generalized additive model in contrast is of the form

$$\mu(X) = E[Y|X_1, \ldots, X_p] = \alpha + f_1(X_1) + \ldots + f_p(X_p)$$

As last time, the f_j's are unspecified smooth "nonparametric" functions.

The approach that is taken to fitting is to fit each function using a scatterplot smoother such as

- Cubic smoothing spline (available in **S** and **SAS**)
- LOESS (available in **S** and **SAS**)
- Kernel smoother (currently not available, but could be added to **S** in theory)
- Thin-plate splines, which allow for interactions between two predictor variables (available in **R** and **SAS**)

When the data is fit, each scatterplot smoother is fit simultaneously using a backfitting algorithm.

In **R** there are two libraries for fitting GAMS, `gam` which matches the **S-Plus** procedures and `mgcv`. `mgcv` takes a different approach to picking the amount of smoothness and the only one that will handle thin-plate splines. In the following examples, `gam` will be used, since it is consistent with **S-Plus** and the output from **SAS**, which will be focused on.

In both packages, actually more general models can be fit. As mentioned, interactions can be added with the use of thin-plate splines. In addition, the semiparametric model can be fit. It has the form

$$E[Y|X, Z_1, \ldots, Z_q] = X\beta + f_1(Z_1) + \ldots + f_q(X_q)$$

In this case, a parametric form is assumed for variables in the vector X. This may contain interactions between the Xs and categorical factors. Nonparametric relationships are assumed for $Z_1, \ldots, Z_q$, which in theory could have some of them be contained in X, though usually they are not. Both **S** and **SAS**can handle this extension.

The final extension to this model, at least that I wish to discuss, is the extension to other distributions. So as in generalized linear models, the distribution of $Y|X$ can be specified. In addition a link function $g(\mu(X))$ can be specified (at least in theory). Thus, for example, we could have the additive logistic model where

$$\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = \alpha + f_1(X_1) + \ldots + f_p(X_p)$$

and $Y|X$ is binomially distributed.

So the most general form of the generalized additive model has

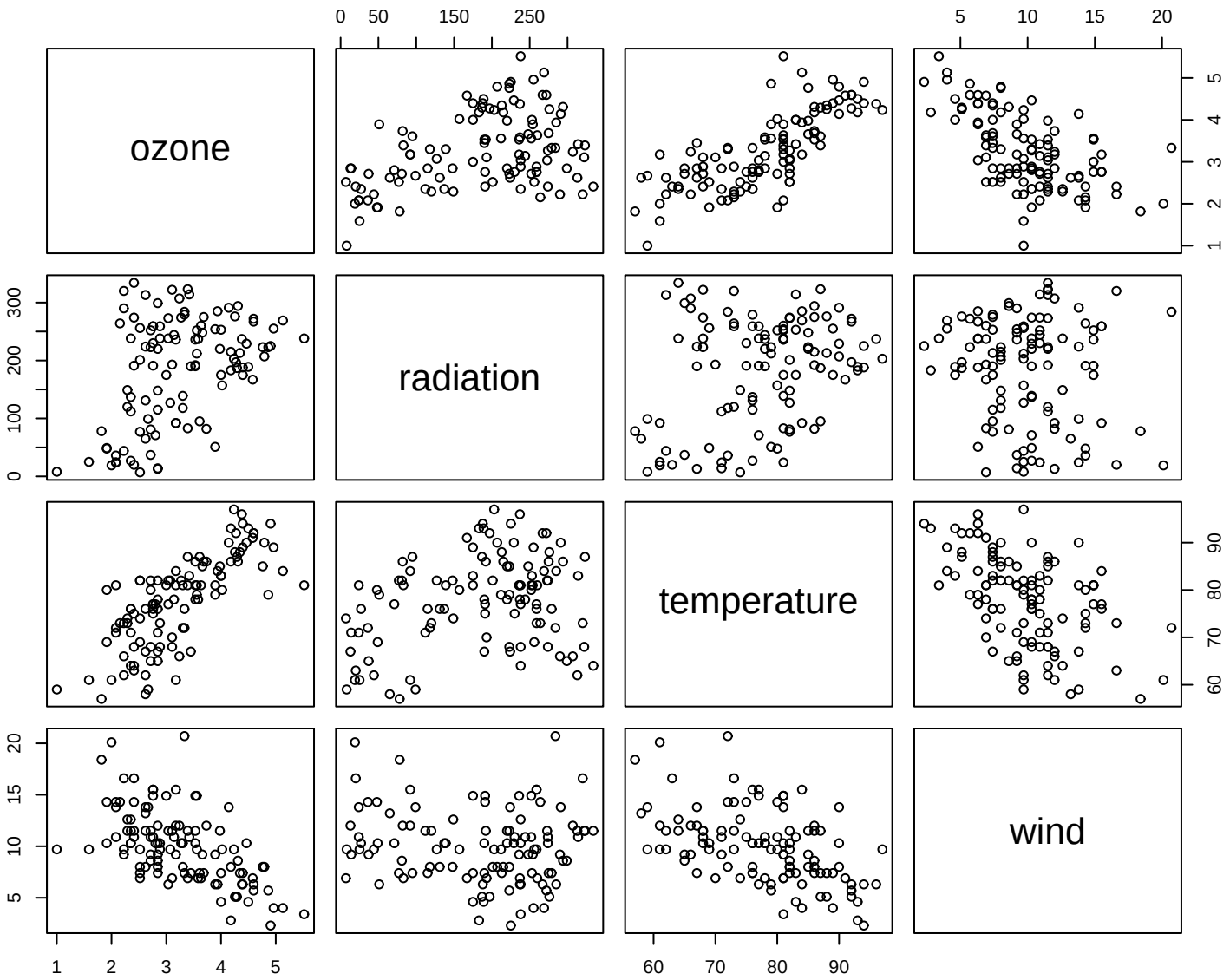
$$g(\mu(X)) = E[Y|X_1, \ldots, X_p] = \alpha + f_1(X_1) + \ldots + f_p(X_p)$$

for some link function $g(\cdot)$. The implementation in **SAS** only allows for the canonical link to be used. In **S**, I believe that any link function available in `glm` can be used in `gam`, which is the function used for describing GAMs.

Example: New York Ozone Concentration

A dataset with 111 observations taken from an environmental study that measured the four variables for 111 consecutive days.

- `ozone`: surface concentration of ozone in New York, in parts per million
- `radiation`: solar radiation
- `temperature`: observed temperature, in degrees Fahrenheit
- `wind`: wind speed, in miles per hour


ozone
radiation
temperature
wind
0 50 150 250
5 10 15 20
1 2 3 4 5
0 100 200 300
60 70 80 90
5 10 15 20

Lets fit this data with **SAS**. The general form of the `PROC GAM` is

```
PROC GAM < option > ;
  CLASS variables ;
  MODEL dependent = < PARAM(effects) > smoothing effects
      < /options > ;
  SCORE data=SAS-data-set out=SAS-data-set ;
  OUTPUT <out=SAS-data-set> keyword < ...keyword> ;
  BY variables ;
  ID variables ;
  FREQ variable ;
```

Most of the options are as we have discussed before. The two important ones are

- `MODEL dependent = < PARAM(effects) > smoothing effects </options >`:

  Describes the model. Model terms can be added by the following possibilities

  - `PARAM(x)`: Adds a linear term in `x` to the model
  - `SPLINE(x)`: Fits smoothing spline with the variable `x`. Degree of smoothness can be controlled by `DF` (default $= 4$).
  - `LOESS(x)`: Fits local regression with variable `x`. Degree of smoothness can be controlled by `DF` (default $= 4$).
  - `SPLINE2(x,y)`: Fits bivariate thin-plate smoothing spline in `x` and `y`. Degree of smoothness can be controlled by `DF` (default $= 4$).

  At least one model term must be defined, though any combination of these is fine.

There are two important options for the `MODEL` statement. They are

- `METHOD = GCV`: Chooses smoothing parameters by generalized cross validation. If `DF` is set for a variable, that choice overrides the `GCV`.
- `DIST = :` Specifies the conditional distribution of the response variable. The choices are `GAUSSIAN` (default), `BINARY`, `BINOMIAL`, `GAMMA`, `IGAUSSIAN`, and `POISSON`. As mentioned before, only the canonical link can be used in each case.

- `SCORE data=SAS-data-set out=SAS-data-set`: Allow for fits to be generated for levels of the predictor variables of interest and stored in **SAS** dataset.

Lets look at the fits of the New York ozone data, generated by the code

```
ods html;
ods graphics on;

PROC GAM DATA = air PLOTS(CLM COMMONAXES);
  MODEL Ozone = SPLINE(Radiation) SPLINE(Wind) SPLINE(Temperature)
  OUTPUT OUT = air_spline PREDICTED RESIDUAL;

RUN;

ods graphics off;
ods html close;
```

Note that the ODS graphics system is an experimental graphics system in **SAS** which allows for generation of the plots that follow. It also generates text output in a number of forms. The example done generates output compatible with web sites (html). This was done as I wanted to get the graphics in gif format so I could convert them to postscript.

The GAM Procedure
Dependent Variable: ozone
Smoothing Model Component(s): spline(radiation) spline(wind)
spline(temperature)

Summary of Input Data Set

Number of Observations	111
Number of Missing Observations	0
Distribution	Gaussian
Link Function	Identity

Iteration Summary and Fit Statistics

Final Number of Backfitting Iterations	7
Final Backfitting Criterion	5.4325795E-9
The Deviance of the Final Estimate	19.942887908

The local score algorithm converged.

Regression Model Analysis

Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.15594	0.49095	0.32	0.7514
Linear(radiation)	0.00235	0.00049382	4.76	<.0001
Linear(wind)	-0.07317	0.01393	-5.25	<.0001
Linear(temperature)	0.04351	0.00540	8.06	<.0001

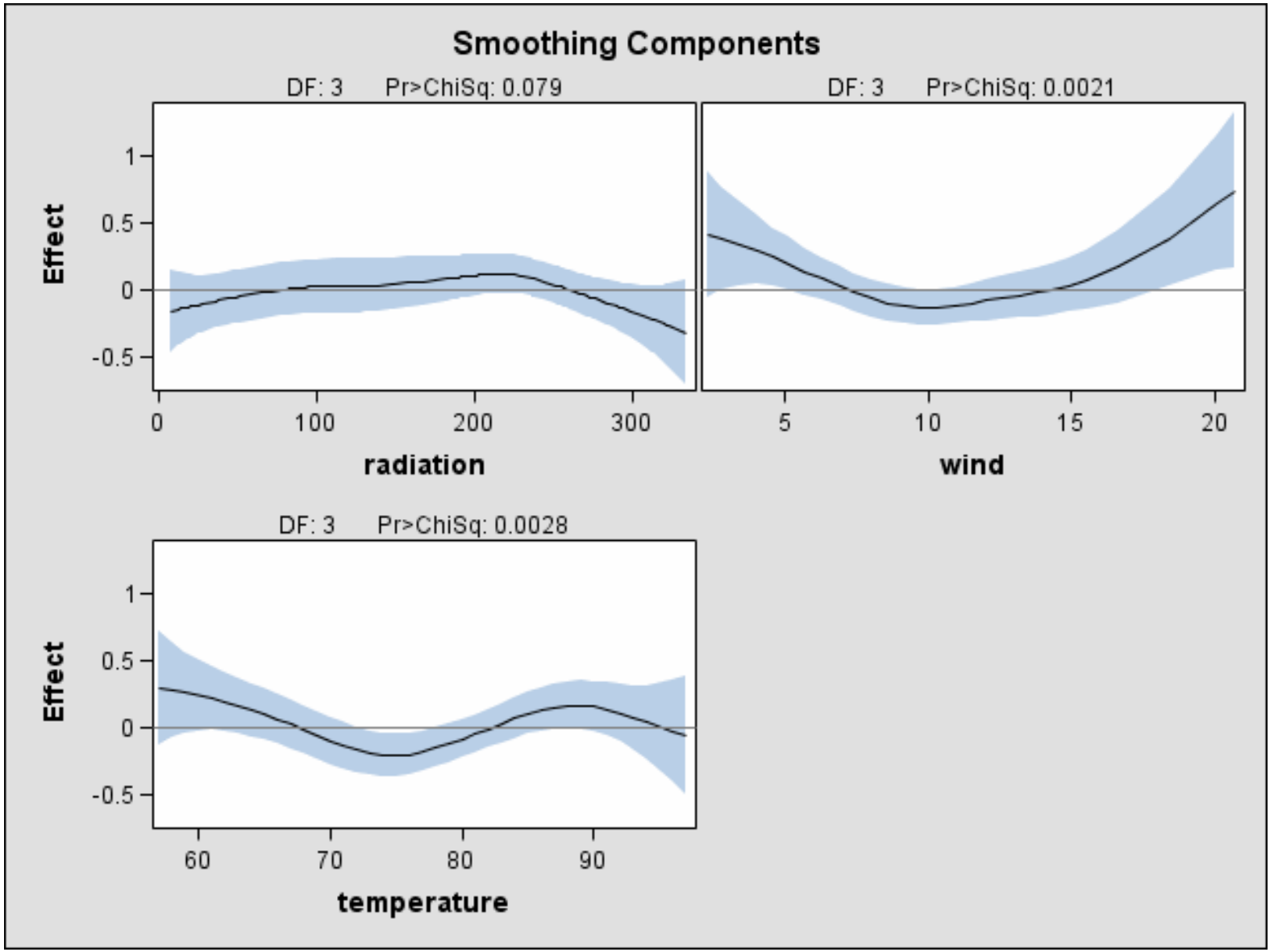
Smoothing Model Analysis

Fit Summary for Smoothing Components

Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(radiation)	0.999893	3.000000	0.182513	93
Spline(wind)	0.991577	3.000000	0.206527	28
Spline(temperature)	0.997312	3.000000	0.157737	39

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(radiation)	3.00000	1.380892	6.7858	0.0790
Spline(wind)	3.00000	2.998911	14.7367	0.0021
Spline(temperature)	3.00000	2.864131	14.0744	0.0028


Smoothing Components
DF: 3 Pr>ChiSq: 0.079
DF: 3 Pr>ChiSq: 0.0021
DF: 3 Pr>ChiSq: 0.0028
Effect
1
0.5
0
-0.5
0
100
200
300
radiation
5
10
15
20
wind
60
70
80
90
temperature

The graphic shows the spline fits of the three effects after the linear trend has been removed.

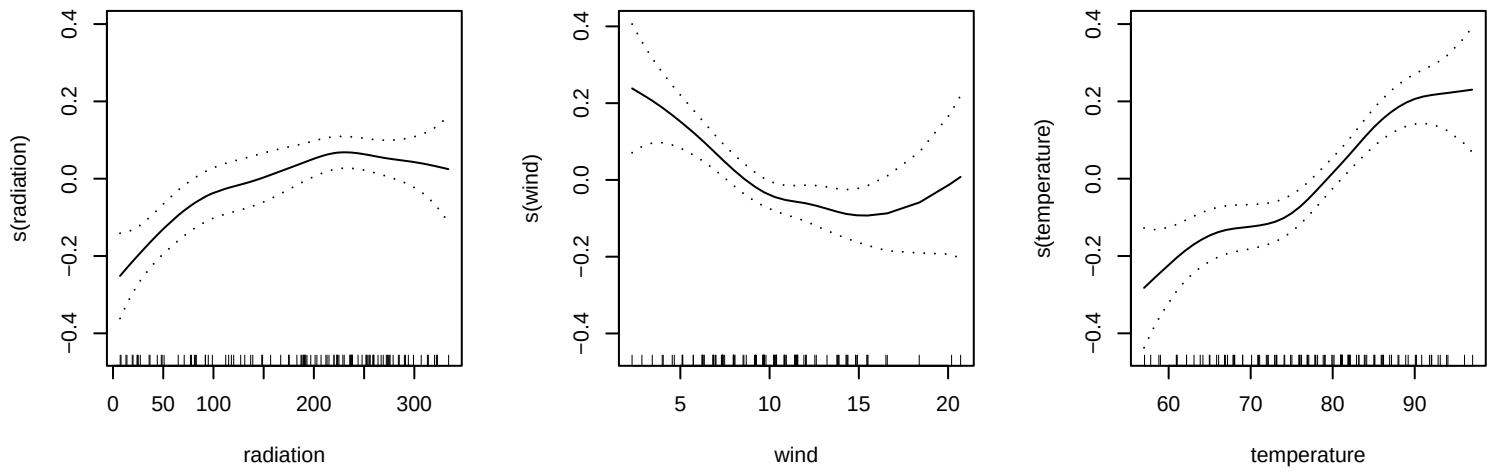
This model can also be fit in **R** will the following code

```
library(gam)  # only needs to be run once per session

# fit linear model for comparison

air.lm <- lm(log(ozone) ~ radiation + wind + temperature,
             data=air)

# fit additive model
air.gam <- gam(log(ozone) ~ s(radiation) + s(wind)
                 + s(temperature), data=air)
summary(air.gam)
anova(air.lm,air.gam)
```

The s function fits a cubic spline for the desired variable. It also takes a df argument with 4 as a default. The output

```
> summary(air.gam)

Call: gam(formula = log(ozone) ~ s(radiation) + s(wind)
        + s(temperature), data = air)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.61526 -0.08185 -0.01468  0.10474  0.33293

(Dispersion Parameter for gaussian family taken to be 0.0248)

    Null Deviance: 9.1633 on 110 degrees of freedom
Residual Deviance: 2.4334 on 98 degrees of freedom
AIC: -81.0436

Number of Local Scoring Iterations: 2
```

```
DF for Terms and F-values for Nonparametric Effects

               Df Npar Df Npar F   Pr(F)
(Intercept)     1
s(radiation)    1       3 2.8496 0.04135 *
s(wind)         1       3 3.4753 0.01893 *
s(temperature)  1       3 2.9350 0.03717 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(air.lm,air.gam)
Analysis of Variance Table

Model 1: log(ozone) ~ radiation + wind + temperature
Model 2: log(ozone) ~ s(radiation) + s(wind) + s(temperature)
  Res.Df     RSS  Df Sum of Sq      F   Pr(>F)
1    107 3.07416
2     98 2.43337   9   0.64079 2.8674 0.004797 **
```




```
par(mfrow = c(1,3), pty="s")
plot(air.gam, se=T, ylim=c(-0.45,0.4))
```

The **S** plots of the fits don't remove the linear trend.

Another example run in **SAS** is

```
PROC GAM DATA = air PLOTS(CLM COMMONAXES);
  MODEL Ozone = SPLINE2(Radiation,Wind) LOESS(Temperature, DF=5);
```

which gives output

```
          Iteration Summary and Fit Statistics

Final Number of Backfitting Iterations                  10
Final Backfitting Criterion                    2.4527292E-9
The Deviance of the Final Estimate             22.048500672


The local score algorithm converged.
```

Regression Model Analysis

Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.24778	0.04435	73.23	<.0001

Smoothing Model Analysis

Fit Summary for Smoothing Components

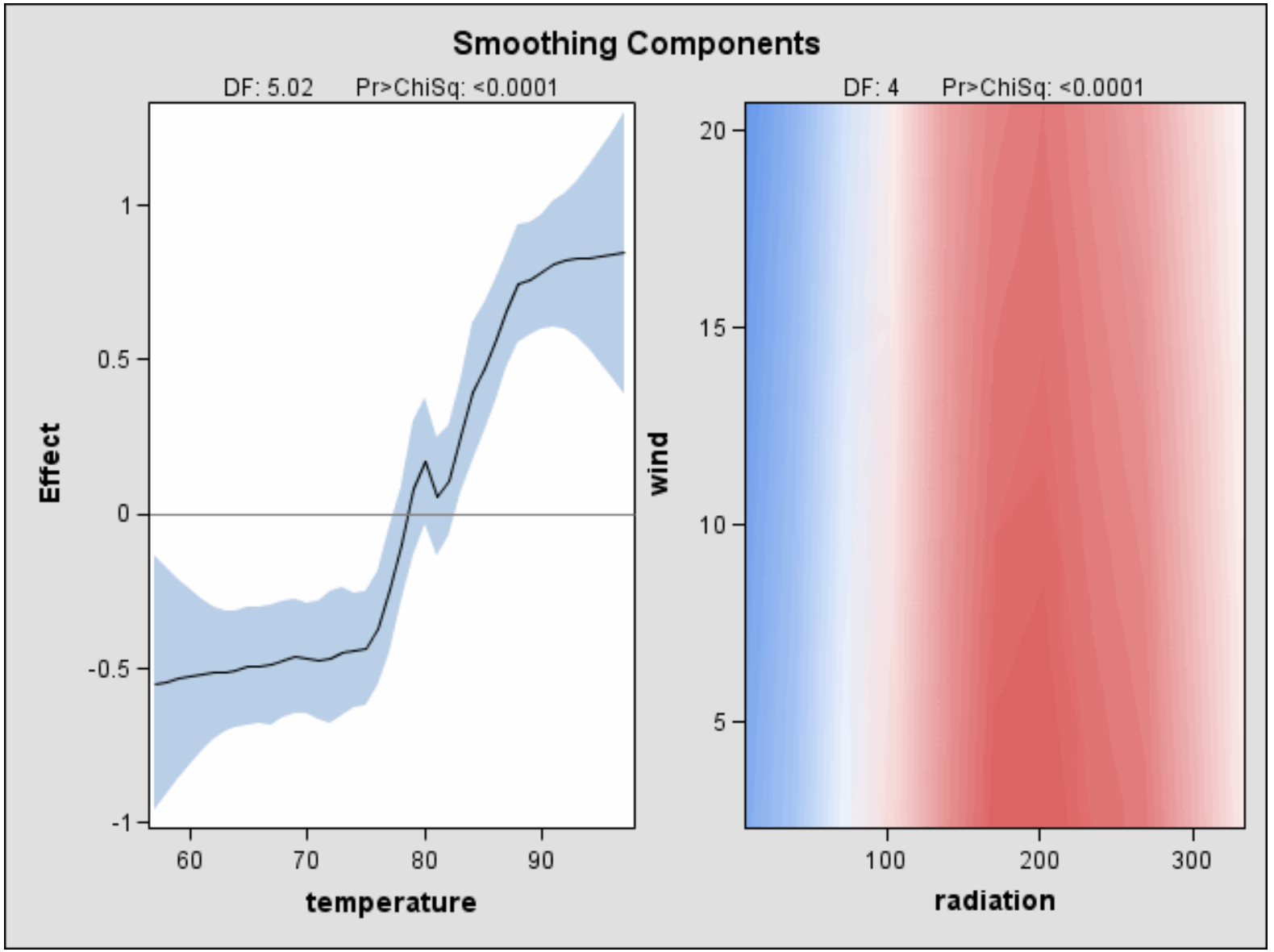
Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline2(radiation wind)	43.311111	4.000000	0.217817	110
Loess(temperature)	0.301802	5.020171	0.002044	111

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline2(radiation wind)	4.00000	9.024662	41.3320	<.0001
Loess(temperature)	5.02017	18.765081	85.9421	<.0001

The following graphic suggests that there probably isn't an interaction since the change in colour is roughly the same for each level of `wind`.

If there was an interaction, I would expect to see a diagonal pattern in the second graph.


Smoothing Components
DF: 5.02
Pr>ChiSq: <0.0001
Effect
1
0.5
0
-0.5
-1
60
70
80
90
temperature
DF: 4
Pr>ChiSq: <0.0001
wind
20
15
10
5
100
200
300
radiation